RICH TRANSCRIPTION 2002 SITE REPORT
PANASONIC SPEECH TECHNOLOGY LABORATORY (PSTL)

Patrick Nguyen, Luca Rigazio, Yvonne Moh, Jean-Claude Junqua

Panasonic Technologies Inc. / Panasonic Speech Technology Laboratory
3888 State Street, Suite 202, Santa Barbara, CA 93105, U.S.A.
email: {nguyen, rigazio, jcj}@research.panasonic.com

ABSTRACT

In this paper, we summarize systems submitted by PSTL to the evaluation. We ran Meta-Data (MD) on Switchboard (SWB) and Broadcast News (BN) data. Speech-to-text systems were built and tested on both SWB and BN systems with limited real-time constraints. For our first participation, our systems were characterized by low complexity, exploratory operating conditions, and small resources.

MD systems served as the segmentation / clustering stage of STT recognizers. Recognizers performed trigram Viterbi decoding with word-internal triphone-models. MLLR adaptation followed optionally.

STT results underlined a relatively favourable benchmark on BN and inauspicious SWB evaluation.

1. INTRODUCTION

The paper will describe components of the systems first. The frontend, acoustic training, segmenter, adaptation, and decoder will be presented. Then, a brief section is devoted to customization of the system. Finally, we give a more general analysis of the results in the history of NIST evaluations.

2. SYSTEM COMPONENTS

The basic building blocks of the system are described here.

2.1. Frontend

The frontend preprocessing is reviewed here.

For BN, only one frontend was used for MD and STT. MFCC parameters were generated as follows. The power spectrum of 32ms is integrated on 20 filter banks. Then, the inverse cosine transform is truncated to 12 cepstral coefficients. The first one is discarded as it models channel gain. A pre-emphasis, high-pass filter is applied. The energy is appended to the cepstrum to obtain 13 coefficients. A five-tap non-causal filter is applied to each parameter to obtain delta coefficients. The same filter is applied and the acceleration coefficients are appended to the output. Then, a purely causal average is computed over 2s and removed from all coefficients. The analysis is then repeated every 10ms. No special processing regarding narrow-band speech was performed.

For SWB/STT, PLP coefficients using an 8-pole prediction, with residual energy were computed. The processing window size was 25ms. We computed a conversation-side mean and variance normalization based on speaker turn segments automatically generated or from the PEM. Delta and acceleration coefficients are computed the same way as above. Variance normalization affected static coefficients only. The feature vector has 27 coefficients.

For SWB/MD, the same PLP analysis was used, but with the sliding-window CMS described above and no variance normalization.

2.2. Acoustic Training

This subsection pertains to acoustic modeling. Most of the training scheme is not original, and we shall therefore put more emphasis on distinctive features.

2.2.1. Broadcast News

Database: LDC97S44 (train96) and LDC98S71 (train97) served as training data. Overlapping speech and music-only segments are discarded. It was cut according to all available tags, including `SyncTime`.

STT-sub10xRT: Word-internal triphonic 3-state HMM models are generated using a decision tree procedure. The training data is split into two equal, arbitrarily chosen, and ideally statistically independent sets. On the first one, a bottom-down decision-tree clustering yields ML decision-tree. Questions might be asked about the phoneme state index, within-word context, and word boundary. To prevent unecessary computations, a minimum of 200 examples is required on each leaf. These trees are then pruned using likelihood evaluated from the held-out set. We enforce

a final size of 2739 states by setting a minimum likelihood change between parent and aggregate children score. For likelihood computations we assume Gaussianity of data.

A classical iterative-splitting training procedure yields 128 Gaussian per state models. The word-segmentation is frozen during that stage. Splitting does not occur for Gaussian seen fewer than 20 times. From the resulting 347832 Gaussians obtained thereby, we retain only 192000. To that end, we adopted a nearest-neighbor, agglomerative clustering scheme. The negative likelihood change, or entropy, functioned as a merging criterion. Merging was not allowed across states. Merging within the entire phoneme not only was extremely expensive, but also did not contribute to recognition.

Gender-dependent are trained using the ML criterion. Variance were fixed to Speaker-Independent (SI) ones. Unseen Gaussians were left untouched (SI), as they seem to provide background modelling necessary for classification. These models provided a basis for a better cut-based segmentation, and the cycle of segmentation, tree growing, iterative splitting, gender-modelling was repeated four times from bootstrap WSJ segmentation.

STT-sub1xRT: The faster-than-realtime models are trained similarly. Merging starts directly at 32 Gaussians per state from 87904 Gaussians to 32000 Gaussians. We allowed merging between all states of allophones with the same center phoneme. The system description incorrectly specifies 87904 component weights. There were only about 40000. We did not train gender-dependent models. The word-level segmentation described above is used to train these models.

MD: Five models, comprising silence, and the cartesian product of bandwidth and gender, were trained on LDC97S44. Wideband was defined as those segments labeled "high fidelity". Music-only detection seemed to increase the false-rejection rate and was discarded.

Gaussian Mixture Models (GMMs) of 512 Gaussians per model segmented LDC98S71 (train96). Iterative splitting and training on both databases yielded our final models.

2.2.2. Switchboard

We used the full 265h of Switchboard1-Release 2 from LDC97S62. Due to lack of time, we made no use of additional material such as CallHome, Cell phone, etc. The segmentation was bootstrapped from a word-level segmentation from Mississippi State University (MSU). All forced-alignments are generated using speaker-adapted models. Cross-word models were trained for all compounds of the SRI language model.

Again, word-internal, 3-state HMM triphones are trained with the same decision-tree clustering and iterative splitting / merging methodology. 3892 mixtures with 128 Gaussians each, totalling 489350 Gaussians were entropy-merged down to 256000 Gaussians.

The segmentation step must decide whether to split compounds or leave them as cross-words. For instance, let `i_have_a` be a trigram compound. It may occur in the training data as `i sil have_a`, `i_have sil a`, or `i sil have sil a`. Its left and right components may also be incorporated with neighbouring compounds. We relied on Viterbi alignement for these decisions.

2.3. Language modeling

2.3.1. Broadcast news

On Table 1 we show which data were incorporated in LM training. The language model includes 53514 words,

Name	Size (M words)	Weight
1996 CSR Hub-4	140	3
North American News	500	1
TDT2 + TDT3	31	3
Acoustic training	1.6	12

Table 1. LM training data: amount and weighting

19007163 bigrams, and 68189884 trigrams in a standard backoff topology.

2.3.2. Switchboard

The language model was kindly provided to us by Andreas Stolcke from SRI [1]. It contains 34610 words, including 1659 compounds, 4826134 bigrams and 11518366 trigrams. Training data includes Broadcast news, Call Home, and Switchboard1. The most frequent compounds were transcribed manually for cross-word coarticulation.

2.3.3. Meta-Data

The meta-data comprises two tasks: speaker segmentation and speaker clustering.

The system used for segmenting the speech in the HUB4 evaluation is as follows. The speech is first decoded using 512-GMM models for the following classes:

- NM : Narrowband Male,
- NF : Narrowband Female,
- WM : Wideband Male,
- WF : Wideband Female,
- sil : Silence.

Each class must last at least 25ms.

The decoded output is then heuristically smoothed in the following rules:

1. Consecutive segments of non-silence are merged, and assigned to the dominating class.

2. Segments surrounded with less than 0.4 seconds are merged again if speech is smaller than 4 seconds.

3. Silence of fewer than 0.15s between two segments of identical conditions are collapsed.

4. Silence of fewer than 0.15s between two segments of different conditions are collapsed, if either segment is less than 4 seconds, a resulting segment encompassing both is labeled with the longer segments' label.

This minimizes the false-rejection (FR), which is later processed with a speech recognizer, but increases the false acceptance (FA) rate. We prefer to overdetect than to chop words.

For speaker clustering, BIC was employed. Bottom-up hierarchial clustering was performed on segments obtained from the segmentation. Statitics used were merged using all data from the belonging to the same cluster. In the first step, each individual segment constitueted its own cluster. Successive merges were performed until a threshold, either by a specific lambda value or a predefined number of clusters, is fulfilled.

3. STT-DECODING

The decoding strategy underwent significant changes during April. To fit the self-imposed real-time constraints, we evaluated many possibilities. Optimizing code and short-listing the vocabulary were the two decisive accelerations that we retained. However, while shrinking the system we realized that we could fit faster-than-real-time constraints, and hence decided to submit it.

In the following we will describe BN-STT and SWB-STT decoding.

3.1. Broadcast news / Less than 10xRT

The BN-STT system proceeds in two stages. The first-pass decoding uses gender-dependent models according to the labels provided by the segmentation/clustering step. There were 192000 Gaussians as described previously.

The most likely transcription is used for MLLR adaptation. Block-diagonal matrices (3 blocks) constitute the transformation. The 7 regression classes were allocated to silence(1), vowels (4), and consonants (2). In degenerate cases we allowed ourselves to reduce the number of classes to three (one for each of silence, vowel, and consonants) or one global class.

Words hypothesized during the first-pass decoding constituted the second-pass lexicon. They amounted to about 400 words per audio cut.

3.2. Broadcast news / Less than 1xRT

The faster sister system also proceeds in two stages. The first-pass decode runs on 2 to 5 seconds of the audio cuts until a minimum amount of true speech is found (about 4 words). Even on that short amount, adaptation provides better recognition and faster decoding time.

Silence Gaussians were not adapted. The statistics of a block-diagonal MLLR transformation were interpolated with the identity matrix prior, using somewhat heuristic weights optimized on a small subset of the eval98 test set.

The second-pass runs with adapted models. The clustering is purely acoustic and does not use any STT results. It is the same as the previous (sub10xRT) system.

3.3. Switchboard / less than 10xRT

Our SWB-STT system is a single-pass Viterbi decoding. We use 256000 Gaussians and trigram language models.

The real-time factor for automatic segmentations leaves almost twice as much time for decoding. Meeting recognition has the same factor but with more channels. RT computation is

$$\mathrm{RT} = \frac{\text{Time spent to produce transcriptions}}{\text{reference time}}. \qquad (1)$$

For reference time, in manual (PEM) evaluation, we count the total speech in the PEM segments. For UEM, however, we count the total audio time in both sides, that is, twice the time of the conversation. Since our segmentation time is very small, we were left with an approximately 10xRT on PEM, and 5xRT on UEM.

In order not to take too much advantage of this speedup, we merely augmented the beam to top-off the real-time constraint. We observed about 0.5-1% WER degradation using automatic labels, and about 1% WER improvement leveraging the large beam.

4. RESULTS

In this section, we compare evaluation versus development results. BN-STT was surprisingly easy, while SWB-STT appeared more difficult.

4.1. BN

Table 2 shows BN-STT system results on our development set (evaluation set of 1998). We excluded training data from the test epoch from the language model. However, a proximity in topic with other TDT data might be the cause of this unexpectedly low WER. We suspect that the LIMSI also observed a similar, although less dramatical "improvement".

Test set	system	WER
Eval98	sub1xRT-first pass	28.4%
	sub1xRT-two pass	$\approx$ 26%
Eval02	sub1xRT	23.7%
Eval98	sub10xRT-first pass	23.4%
	sub10xRT-two pass	21.7%
Eval02	sub10xRT	20.1%

Table 2. BN-STT system

4.2. SWB

Results for Switchboard-I only on Table 3. Due to lack of resources, we had been mainly developing and testing Switchboard-I data. Our development set, SWBD-00, consists of the Switchboard-I test set of the NIST 2000 evaluation (also the dev set of the 2001 eval). We include our system on the dev set, our system on Eval02, AT&T's faster-than-realtime system, and reported results by JHU in 2001, with a single-pass running at about 25xRT, if we use the lenient unpartitioned evaluation RT computation. CU-HTk submitted a "late" system after the evaluation, which we list in the table. It was developed in a very short time.

Test set	WER	RT
SWBD-00	33.8%	10
Eval02	36.7%	10
JHU.1	32.7%	50
AT&T-1x	29.5%	1
CUHTk-late	22.3%	10

Table 3. SWB-STT system

The relative drop in performance between dev and eval sets could be due to overtuning or instrinsic difficulty. Our lack of experience in participating to these evaluations did also contribute.

Real-time factors extracted from JHU's presentation are also in line with other participants of that evaluation. We used them to calibrate our expectations on our relative position in the evaluation. In our projections, our system was to be comparable with a standard first-pass of eval-2001, in half the real-time factor.

Nonetheless, we hope to have aroused interest among other participants.

5. PERFORMANCE COMPARISON

This section is devoted to trying to interpret our relatively low results. Given the very high technical level of NIST evaluations, first-time participants such as PSTL customarily display relatively high WER compared with old-timers.

Tables 4 and 5 show approximate best and worst participants scores for BN and SWBD-I data along the years. Results are not directly comparative from year to year, especially for SWB data, to which recognizers seem to be very sensitive.

	96	97	98	99	02
Best	27%	16%	14%	14%	13%
Worst	56%	39%	26%	28%	20%
Participants	10	10	10	4	2

Table 4. WER of BN-STT

	<99	00	01	02
Best	37%	19%	20%	22%
Worst	?%	42%	37%	37%
Participants	?	5	7	7

Table 5. WER of SWB-STT (Switchboard-I)

For BN data, we seem to score somewhat decently as a first-timer. On SWB, however, the outlook is not overly alarming but in the grey area. Even with our lower than expected performance, we score relatively better.

6. CONCLUSION AND OUTLOOK

In this paper, we presented our MD and STT systems submitted to the RT-2002 evaluation. Most of characteristics are but standard ones, yet we attempted to focus on exploratory conditions, which are characterized by small resources. Indeed, except for the standard 10xRT and the late HTk system, we were the only ones in the sub10xRT UEM SWB and sub1xRT BN, and MD conditions. Rather to concentrate on a particular condition, we decided to evaluate what could be done in non-standard conditions. Also, our system is portable in the sense that almost no SWB- or BN-specific customizations were considered. We were the most cooperative in submitting as many primary systems possible in STT and MD.

We will continue implementing baseline features towards reaching an acceptable level of performance. Our unreleased distinctive features, such as eigenvoices, constrained model space adaptation, full-variance LU MAP adaptation, etc. are being currently tested.

7. REFERENCES

[1] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. Ramana Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 Conversational Speech Transcription System," in *Proc. of 2000 Speech Transcription Workshop*, 2000.

[2] P. Nguyen, L. Rigazio, C. Wellekens, and J.-C. Junqua, "Construction of model space constratins," in *Proc. of ASRU-2001*, 2001.

[3] P. Nguyen, L. Rigazio, C. Wellekens, and J.-C. Junqua, "LU factorization for feature transformation," in *Submitted to ICSLP-2002*, 2002.

[4] P. Nguyen, L. Rigazio, and J.-C. Junqua, "EWAVES: an efficient decoding algorithm for lexical tree based speech recognition," in *Proc. of ICSLP*, Beijing, China, Oct. 2000, vol. 4, pp. 286–289.